Homework 1

Due: 23:59, --

---

1. **New York Times** (50 points, Adapted from Rachel Schutt at Columbia) There are 31 datasets named nyt1.csv, nyt2.csv,...,nyt31.csv, which can be found on my webpage. Each one represents one day's worth of ad impressions and clicks on the New York Times homepage in May, 2012 (these are simulated). Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number of impressions (page views), number clicks (actions) and whether the user was logged.in.

WARNING: These data are designed to simulate real data (and all the accompanying headaches). Don't make assumptions (e.g., that age > 0) unless you've verified that the data actually comply with these assumptions.

   a. Create a new variable, age_group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64" and "65+".
   b. For a single day
      i. Plot the distributions of number impressions and click-through-rate (CTR=# clicks/# impressions), for these 6 age categories. [You will turn in a .R and .html file where the latter will show this plot]
      ii. Define a new variable to segment or categorize users based on their click behavior

      (The remaining questions require you to be creative and come with you your own variables/parameters/metrics are required in the question.)

      iii. Explore the data and make visual and quantitative comparisons across user segments/ demographics (<18 year old male vs < 18 year old females or logged-in vs not, for example).
   c. Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time; what will compress the data, but still capture user behavior. Now extend your analysis across days (one week is sufficient). Visualize metrics and distributions over time. Your plot should emphasize what actually changes over days.
   d. Describe and interpret any patterns you find. [Include in your writeup]
2. **Your Data** (50 points) For the second part of this assignment, you need to find some data of your own. After you've found the data that you plan to use, post about it on Piazza (also explain how you found it). I want everybody to have **different** data, so

posting about it on ELMS will make it **off limits** to everybody else (this is an incentive to get it done early). The data do not have to be publicly accessible (e.g. you can use personal / professional data), but you should have every right to distribute and discuss the data (don't do anything sketchy to get the data).

a. Thoroughly describe the data using the analysis and visualization techniques we covered in class (but feel free to go beyond them). I should come away with an understanding of your data. A good dataset will:
    i. Not be too small (If you can understand everything that's going on by looking at it, it's a bad dataset). I'll stipulate that the dataset must at least have 250 observations (but it can have much more!).
    ii. Be "real" - don't make it up or use a purpose-built computer program to generate it
b. Explain what (if anything) you had to do to make the data usable. This will likely be influenced by what visualizations analyses you want to do with your data.
c. Create 2-3 slides that summarizes your work, to be presented in class the next week for "Lightning Talks".

Turn it In

Turn in your code by submitting via ELMS.